Theory of mind on robots: A fMRI Study

Florian Niefind, SS 10

Seminar: Multimodal Interaction with intelligent agents,
Lecturers: Crocker, Staudte

The study

- Subjects play a variant of the Prisoner's Dilemma Game (PDG) against:
  - another human
  - an anthropomorphic robot
  - a functional robot
  - a computer program
- At least that's what they think they do...
- BOLD is being measured via fMRI

PDG

- players either cooperate or defect
- based on responses of both, rewards are assigned
- the goal is to get as many points as possible (and more than the opponent)

C-P1	C-P2	R-P2	R-P2
0	0	0	0
0	1	10	20
1	0	20	10
1	1	20	20

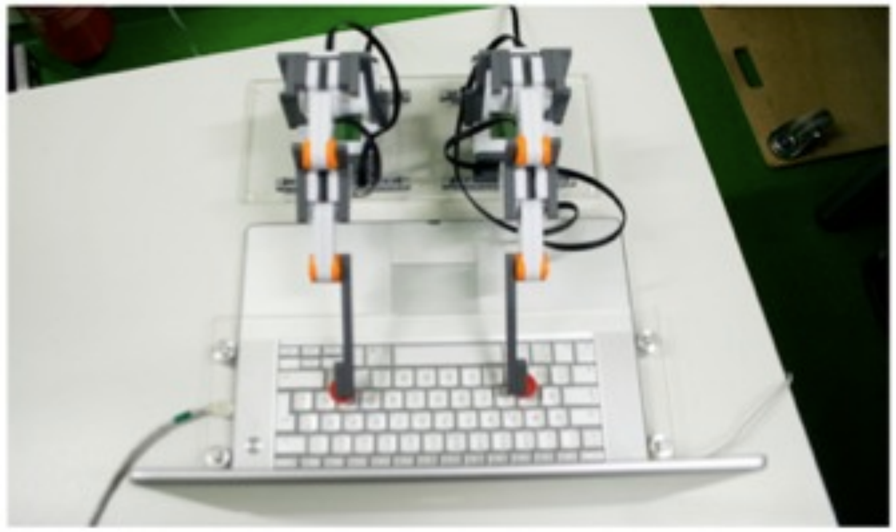
Research Interest

- PDG gives a way to investigate ToM related brain areas
  - responses are randomized
  - thus only the effect of expectations (ToM) can be measured
- test influences of appearance/ embodiment of a robot on ascribed intentionality

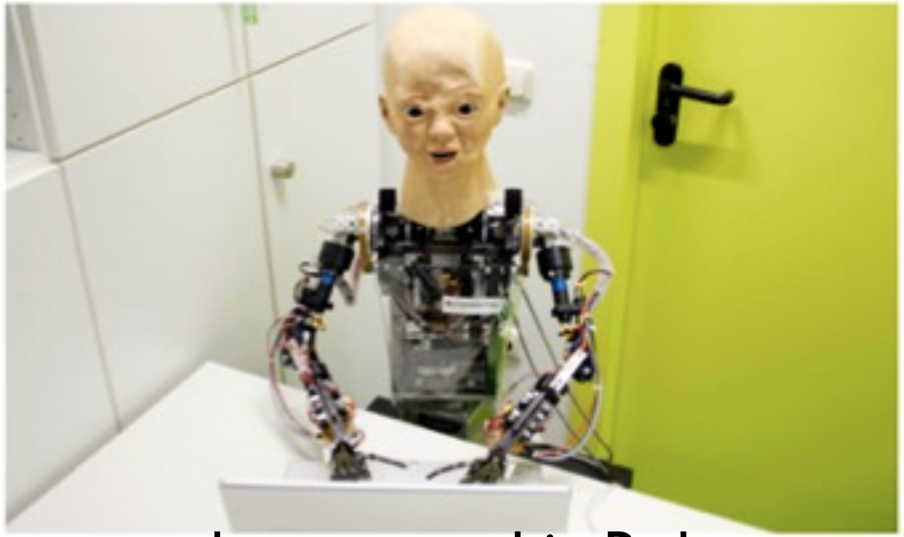
Setting

- Participants were briefed on the game in the presence of the four opponents
- then were put into the MR-Scanner and played ten blocks of 9 single games in each condition plus 1 baseline condition

Pictures

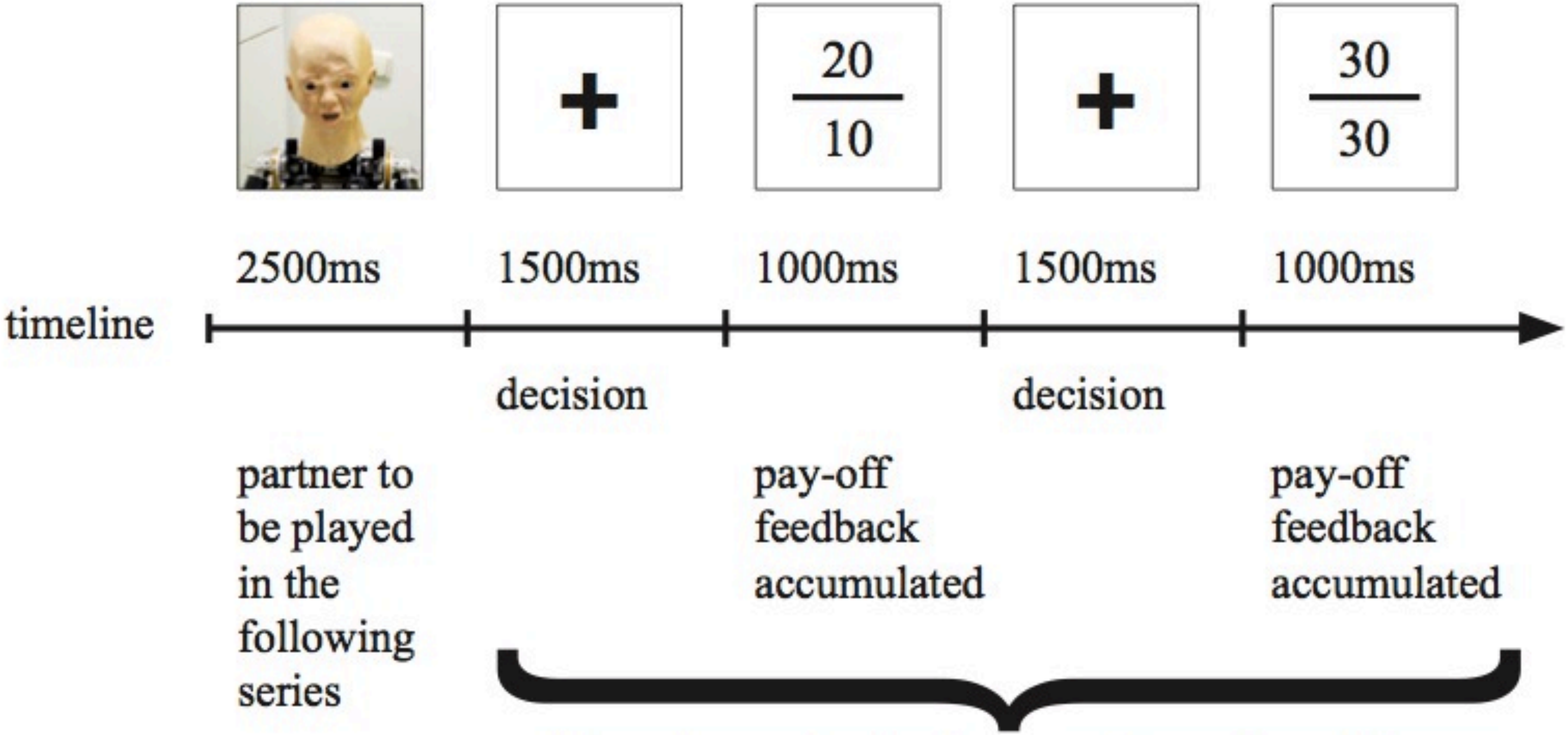


functional Robot



anthropomorphic Robot

Briefing



20 / 10

30 / 30

2500ms 1500ms 1000ms 1500ms 1000ms

timeline

decision

decision

partner to be played in the following series

pay-off feedback accumulated

pay-off feedback accumulated

Ten series a nine single games for each condition (HP, AR, FR, CP, or baseline)

Behavioral Results

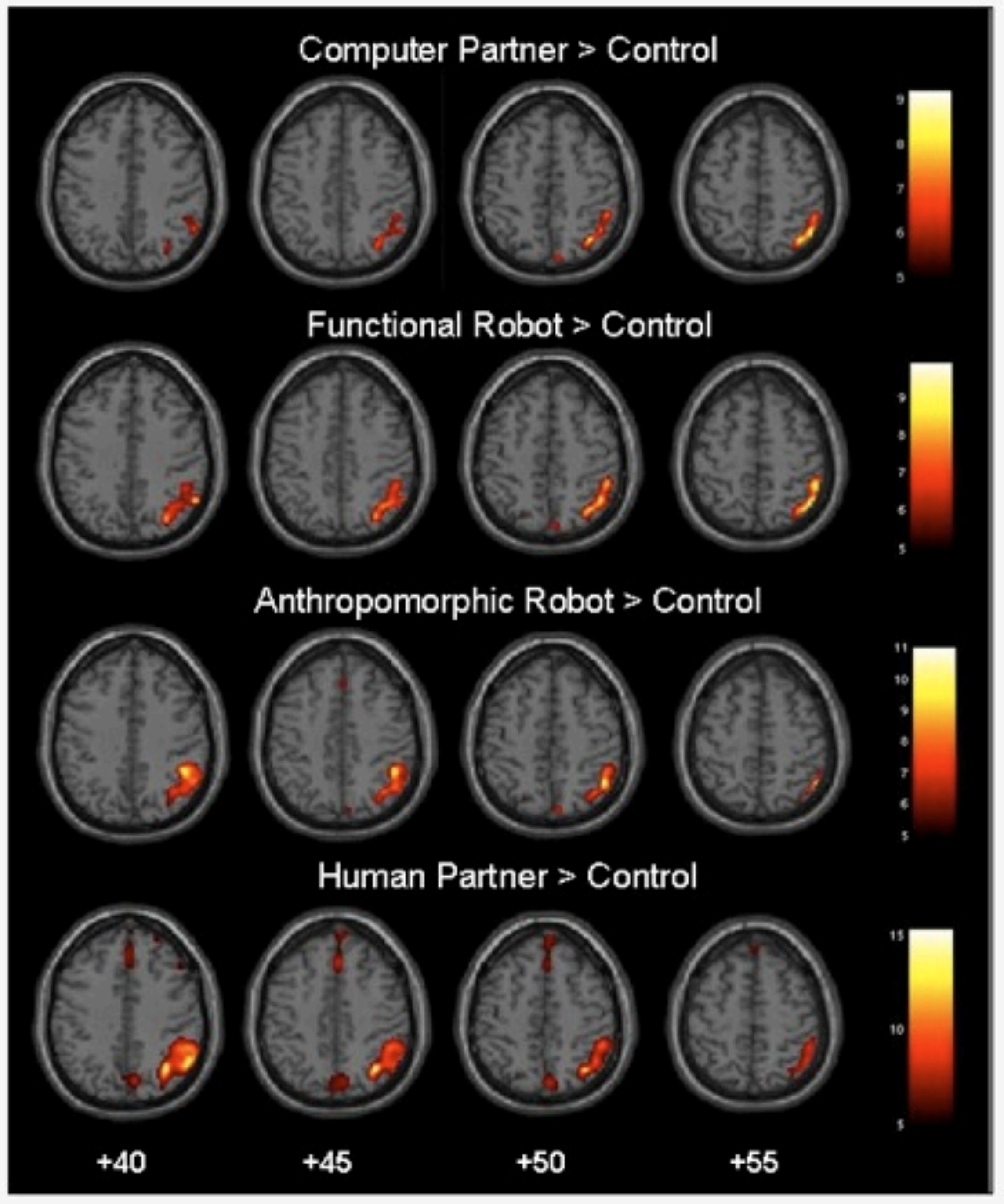
- participants played rather competitive in all 4 conditons (60/40, competitive/cooperative)
- participants rated the opponent as more intelligent and fun to play with the more human-like they were in appearance
- the human-like robot and the human were rated more friendly and sympathetic then the functional robot, but also more competitive

Neuroimaging Results

- activation in temporo-parietal junction increased with perceived human-likeness

- medial prefrontal cortex activation only for human-like opponents

- the results exhibit a significant linear trend


Computer Partner > Control
Functional Robot > Control
Anthropomorphic Robot > Control
Human Partner > Control
+40
+45
+50
+55

Interpretation

- method works
- participants ascribe intentionality to all interactors
- effect for human stronger then for all other conditions
- three out of four empathized more with the robots then with the computer program

Relations to our seminar

- new methodology to show that people assign intentions to partners
- to quantify how much?
- study investigated influences of appearance of a robot
    - other factors can be investigated as well

However

- general problem of neuroimaging:
  - we don't know what exactly is going on in the nicely glowing areas
  - BOLD linking hypothesis still doubted
- baseline was not carefully designed: no planning involved here

General Issues

- functions of gaze, gestures, appearance/motion for communication
    - structural organisation, information relevance, encoding additional information, triggering ‘Like-me’ hypothesis
    - to a certain extent any modality can be used for any of these functions
    - some are better for certain tasks: high expressivity of language
- of course helpful in HRI as long as they look natural enough to be processed automatically instead of taking away rescources

Joint attention in HRI

- from the theoretical viewpoint there is no argument that principally speaks against it
  - as intentionality is involved in joint attention the whole debate can be transferred to philosophy of mind
  - theoretical positions like functionalism allow for intentionality to arise from any physical system
- the question thus is an empirical one:
  - judging from the work we have seen, empirical research on the prerequisites of joint attention is still in its early stages
  - no strong case against the possibility of joint attention